대 한 민 국 특 허 청
KOREAN INDUSTRIAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Industrial Property Office.

출 원 번 호 : 특허출원 2000년 제 58759 호
Application Number

출 원 년 월 일 : 2000년 10월 06일
Date of Application

출 원 인 : 삼성전자 주식회사 외 1명
Applicant(s)

2000 년 12 월 23 일

특 허 청
COMMISSIONER

CERTIFIED COPY OF PRIORITY DOCUMENT

【서류명】 특허출원서
【권리구분】 특허
【수신처】 특허청장
【참조번호】 0011
【제출일자】 2000.10.06
【국제특허분류】 H04N
【발명의 명칭】 특징 벡터 데이터 공간의 인덱싱 방법
【발명의 영문명칭】 Indexing method of feature vector data space
【출원인】
【명칭】 삼성전자 주식회사
【출원인코드】 1-1998-104271-3
【출원인】
【명칭】 더 리전트 오브 더 유니버시티 오브 캘리포니아
【출원인코드】 5-1999-020685-1
【대리인】
【성명】 이영필
【대리인코드】 9-1998-000334-6
【포괄위임등록번호】 1999-009556-9
【포괄위임등록번호】 2000-055014-6
【발명자】
【성명의 국문표기】 최양림
【성명의 영문표기】 CHOI,Yang Lim
【주민등록번호】 710120-1830615
【우편번호】 463-060
【주소】 경기도 성남시 분당구 이매동 124 이매 한신아파트 210동 1509호
【국적】 KR
【발명자】
【성명의 국문표기】 비 .에스.만주나스
【성명의 영문표기】 B.S.,Manjunath
【주소】 미합중국 캘리포니아 93106-9560 산타바바라, 유니버시티 오브 캘리 포티아
【국적】 US

【발명자】

【성명의 국문표기】	펭 우
【성명의 영문표기】	PENG,Wu
【주소】	미합중국 캘리포니아 93106-9560 산타바바라, 유니버시티 오브 캘리 포니아
【국적】	US

【우선권주장】

【출원국명】	US
【출원종류】	특허
【출원번호】	60/226,586
【출원일자】	2000.08.21
【증명서류】	첨부

【심사청구】 청구

【취지】 특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인
이영필 (인)

【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	6	면	6,000 원
【우선권주장료】	1	건	26,000 원
【심사청구료】	14	항	557,000 원

【합계】 618,000 원

【첨부서류】 1. 요약서·명세서(도면)_1통 2.우선권증명서류 및 동 번역문_1통[추후제출]

【요약서】

【요약】

특징 벡터 데이터 공간내에서 특징 벡터를 인덱싱하는 방법이 개시된다. 본 특징 벡터 데이터 공간의 인덱싱 방법은 (a) 특징 벡터 데이터 공간내의 특징 벡터 데이터의 통계적 분포를 기초로 특징 벡터들의 근사화를 적응적으로 구성하는 단계를 포함하는 것을 특징으로 한다. 본 발명에 의한 특징 벡터 데이터 공간의 인덱싱 방법은 일반적으로 특징 벡터들이 균일하게 분포하지 않는 차수(dimensionality)가 높은 벡터 공간내에서 효율적으로 인덱싱한다. 또한, 상기와 같은 특징 벡터 데이터 공간의 인덱싱 방법은 새로운 특징 벡터 데이터가 추가되었을때 인덱싱의 업그레이드가 용이하다는 장점이 있다.

【대표도】

도 1

【명세서】

【발명의 명칭】

특징 벡터 데이터 공간의 인덱싱 방법{Indexing method of feature vector data space}

【도면의 간단한 설명】

도 1은 본 발명의 실시예에 따른 인덱싱 방법의 주요 단계들을 나타낸 흐름도이다.

도 2는 각 차원상에서 데이터의 주변 분포가 균등하다고 하더라도 데이터의 결합 분포(joint distribution)는 여전히 균등하지 않고 응집되어 있는 경우를 설명하기 위한 도면이다.

도 3a는 특징 벡터 데이터 공간내의 특징 벡터 데이터의 분포를 나타내는 히스토그램이다.

도 3b는 상기 히스토그램에 대한 추정된 확률 분포 함수를 나타낸 그래프이다.

도 4a는 데이터 집합들의 특징 벡터값들을 나타낸 그래프이다.

도 4b는 도 4a의 데이터 집합에 대한 히스토그램의 계산 결과를 나타낸 그래프이다.

도 4c, 도 4d, 및 도 4e는 추정에 사용된 엘리먼트들의 양이 각각 1700, 3400, 및 5000일 때의 추정된 확률 분포 함수를 나타낸 그래프이다.

도 5a 및 도 5b는 종래의 인덱싱 방법과 본 발명의 인덱싱 방법을 사용하여 제1 단계 필터링 및 제2 단계 필터링에서 방문한 특징 벡터의 수를 비교 도시한 그래프이다.

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<9> 본 발명은 특징 벡터 데이터 공간의 인덱싱 방법에 관한 것으로, 더 상세하게는 특징 벡터들이 균일하게 분포하지 않는 차수(dimensionality)가 높은 벡터 공간내에서 효율적으로 인덱싱을 수행하는 특징 벡터 데이터 공간의 인덱싱 방법에 관한 것이다.

<10> 일반적인 멀티미디어 데이터 기술자(multimedia data descriptors)는 고차원성(high dimensionality)을 가지기 때문에 효율적인 인덱싱 스킴을 설계하는데 있어 장애가 되고 있다. 따라서, 최근에는 새로운 인덱스 구조가 제안되고 있다. 이러한 인덱스 구조들은 공통적으로 벡터 공간내에 특징 벡터 데이터들이 균일하게 분포하고 있다고 가정하고 있다. 하지만, 영상 텍스쳐 기술자들과 같은 많은 매체 기술자들은 균일하게 분포하지 않는다. 예를들어, 잘 알려진 VA(vector approximation: 벡터 근사화) 파일을 사용하는 방법의 경우, 그 방법의 성능은 특징 벡터의 균일도에 의존하고, 일반적으로 특징 벡터들이 균일하게 분포하지 않는 차수(dimensionality)가 높은 벡터 공간내에서 특징 벡터 데이터들을 인덱싱할 때 성능이 현저하게 떨어진다는 문제점이 있다.

【발명이 이루고자 하는 기술적 과제】

<11> 본 발명이 이루고자 하는 기술적 과제는 특징 벡터들이 균일하게 분포하지 않는 차수가 높은 벡터 공간내에서 효율적으로 인덱싱을 수행하는 특징 벡터 데이터 공간의 인덱싱 방법을 제공하는 것이다.

【발명의 구성 및 작용】

<12> 상기 과제를 이루기 위하여 본 발명에 따른 특징 벡터 데이터 공간의 인덱싱 방법은 (a) 특징 벡터 데이터 공간내의 특징 벡터 데이터의 통계적 분포를 기초로 특징 벡터들을 적응적으로 근사화함으로써 특징 벡터 데이터 공간을 인덱싱하는 단계;를 포함하는 것을 특징으로 한다.

<13> 또한, 상기 (a) 단계는, (a-1) 특징 벡터 데이터 공간내에서 특징 벡터 데이터의 통계적인 분포를 측정하는 단계; (a-2) 상기 통계적인 분포를 사용하여 데이터의 주변 분포를 추정하는 단계; (a-3) 상기 추정된 분포를 그 데이터가 어떠한 각 그리드에 놓이는 확률이 균등하게 되는 복수 개의 그리드들로 분할하는 단계; 및 (a-4) 분할된 그리드들을 사용하여 특징 벡터 데이터 공간을 인덱싱하는 단계;를 포함하는 것을 특징으로 한다.

<14> 또한, 상기 (a-4) 단계 이전에, 새로운 데이터가 들어오면 이전 확률 분포 함수와 갱신된 확률 분포 함수를 기초로 그리드를 갱신하는 단계;를 더 포함하는 것이 바람직하다.

<15> 또한, 상기 (a-4) 단계는, VA(vector approximation: 벡터 근사화) 파일을 사용하여 인덱싱하는 단계;를 포함하는 것이 바람직하다.

<16> 또한, 상기 복수 개의 그리드의 수는 그 차원에 할당된 주어진 비트수에 의하여 결정되는 것이 바람직하다.

<17> 또한, 상기 (a-2) 단계는, (a-2-1) 확률 분포 함수를 소정의 분포 함수의 가중된 합을 사용하여 정의하는 단계; 및 (a-2-2) 상기 (a-2-1) 단계에서 정의된 확률 분포 함

수를 사용하여 소정의 변수들을 추정함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것이 바람직하다.

<18> 또한, 상기 (a-2-2) 단계는, 상기 (a-2-1) 단계에서 정의된 확률 분포 함수를 사용하여 최대화 추정 알고리즘을 기초로, N은 소정의 양의 수, 매번 주어진 모든 N 데이터를 사용하여 반복적으로 소정의 변수들을 추정함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것이 바람직하다.

<19> 또한, 상기 소정의 분포 함수는 가우시안 함수인 것이 바람직하다.

<20> 또한, 상기 (a-2-1) 단계는 $p(x|j)$는 $p(x|j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\}$로서 정의되고, 혼합 변수라고 칭해지는 계수 $P(j)$는 $0\le P(j)\le 1$의 조건과, $\sum_{j=1}^{M}P(j)=1$의 조건을 만족한다고 할 때, 일차원 신호의 확률 분포 함수 $p(x)$를 $p(x)=\sum_{j=1}^{N}p(x|j)P(j)$라고 가정하는 단계;를 포함하는 것이 바람직하다.

<21> 또한, 상기 (a-2-2) 단계는, 추정하여야 하는 변수들은 $j=1,\ldots,M$이고, $l=1,\ldots,N$이며, $v[l]$는 주어진 데이터 세트라 할 때, $\Phi(\phi_1,\ldots,\phi_M)=\prod_{l=0}^{N}p(v[l]|(\phi_1,\ldots,\phi_M))$를 최대로 하는 Φ를 구함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것이 바람직하다.

<22> 또한, 상기 (a-2-2) 단계는, t는 반복 횟수를 나타내는 양의 수라 할 때, $\mu_j^{t+1}=\frac{\sum_{l=1}^{N}p(j|v[l])^t v[l]}{\sum_{l=1}^{N}p(j|v[l])^t}$, $(\sigma_j^2)^{t+1}=\frac{\sum_{l=1}^{N}p(j|v[l])^t(v[l]-\mu_j^t)^2}{\sum_{l=1}^{N}p(j|v[l])^t}$, $P(j)^{t+1}=\frac{1}{N}\sum_{l=1}^{N}p(j|v[l])^t$에 따라 추정된 변수들을 구하는 단계;를 더 포함하는 것이 바람직하다.

<23> 또한, 상기 (a-2-2) 단계는, N 데이터

$v[I]$을 사용하여 추정된 변수 집합 $\{P(j)^N, \mu_j^N, (\sigma_j^2)^N\}$이 주어질 때, 새로운 데이터 $v[N+1]$가 들어오면 $\mu_j^{N+1}=\mu_j^N+\theta_j^{N+1}(v[N+1]-\mu_j^N)$, $(\sigma_j^2)^{N+1}=(\sigma_j^2)^N+\theta_j^{N+1}[(v[N+1]-\mu_j^N)^2-(\sigma_j^2)^N]$, $P(j)^{N+1}=P(j)^N+\frac{1}{N+1}(p(j \mid v[N+1])-P(j)^N)$, 및 $(\theta_j^{N+1})^{-1}=\frac{p(j \mid v[N])}{p(j \mid v[N+1])}(\theta_j^N)^{-1}+1$을 계산함으로써 갱신된 변수 집합을 구하는 단계;를 더 포함하는 것이 바람직하다.

<24> 또한, 상기 (a-2-2) 단계는, 이전 확률 분포 함수와 갱신된 확률 분포 함수를 각각 $\hat{p}_{old}(x)$과 $\hat{p}_{new}(x)$라 할 때, 각 차원에 대하여 $\rho=\frac{\int(\hat{p}_{old}(x)-\hat{p}_{new}(x))^2dx}{\int\hat{p}_{old}(x)^2dx}$으로 정의된 확률 분포 함수의 변화를 측정하는 단계; 및 ρ가 소정의 임계값보다 크면 그 차원에 대하여 근사화를 갱신하는 단계;를 더 포함하는 것이 바람직하다.

<25> 또한, 상기 (a-3) 단계는 추정된 확률 분포 함수를 $\hat{p}(x)$라 할 때, $\int_{c[l]}^{c[l+1]}\hat{p}(x)dx=\frac{1}{2^b}\int_{c[0]}^{c[2^n]}\hat{p}(x)dx$의 조건(criterion)을 만족하는 경계점들 $c[l]$에 의하여 결정되는 복수 개의 그리드들을 사용하여 각 그리드에 의하여 커버되는 면적이 동일하도록 확률 분포 함수를 분할하는 단계;를 포함하는 것이 바람직하다.

<26> 이하 첨부된 도면들을 참조하여 본 발명의 바람직한 실시예들을 상세히 설명하기로 한다.

<27> 도 1에는 본 발명의 실시예에 따른 인덱싱 방법의 주요 단계들을 흐름도로써 나타내었다. 본 발명에 따르면, 특징 벡터 데이터 공간내의 특징 벡터 데이터의 통계적 분포를 기초로 VA 파일을 적응적으로 구성한다. 즉, 밀하게 분포하고 있는 셀들은 인덱싱 성능을 열화시킬 수 있으므로 본 발명에서는 데이터의 통계적 특성에 따라 특징 벡터들의 근사화를 적응적으로 구성한다.이를 위하여, 본 발명에 따른 인덱싱 방법에서는, 먼저, 특징 벡터 데이터 공간내에서 특징 벡터 데이터의 통계적인 분포를 측정한다(단계 102).

다음으로, 상기 통계적인 분포를 사용하여 데이터의 주변 분포(marginal distribution)를 추정한다(단계 104). 다음으로, 추정된 분포를 그 데이터가 어떠한 각 그리드에 놓이는 확률이 균등하게 되는 복수 개의 그리드들로 분할한다(단계 106). 그리드의 수는 그 차원에 할당된 주어진 비트수에 의하여 결정된다. 이제, 분할된 그리드들을 사용하여 특징 벡터 데이터 공간을 인덱싱한다(단계 108). 단계(108)는 잘 알려진 VA(vector approximation: 벡터 근사화) 파일을 사용한 인덱싱 방법을 기초로 하는 것이 가능하다.

<28> 상기와 같은 방법으로 구성된 근사화는 밀하게 분포하는 셀들을 가질 가능성을 저감시킨다. 따라서, 인덱싱 성능이 향상된다.

<29> 여기서, 데이터의 주변 분포는 데이터의 고차원 분포의 공간적 정보만을 캡처(capture)한다는 것에 주목하여야 한다. 도 2에는 각 차원상에서 데이터의 주변 분포가 균등하다고 하더라도 데이터의 결합 분포(joint distribution)는 여전히 균등하지 않고 응집되어 있는 경우를 설명하기 위한 도면을 나타내었다. 도 2를 참조하면, 전체 특징 벡터 데이터 공간(20)내의 각 차원상에서 데이터의 주변 분포는 균등하다. 하지만, 다른 차원들의 데이터와의 결합 분포(joint distribution)는 여전히 균등하지 않고 응집되어 있다. 하지만, 차원(dimensionality) 및 영상/비디오 데이터베이스의 속성을 증가시킴에 따라 다른 차원들 상에서 데이터의 상관은 줄어든다는 것을 고려하면 고차원 데이터의 통계적 특성을 캡처함으로써 주변 분포를 예측하는 것은 여전히 효율적인 방법될 수 있다.

<30> 이하에서는 상기와 같은 본 발명의 개념을 구현하기 위한 방법들을 보다 상세하게 설명한다. 먼저, 차원

i상의 데이터의 확률 분포 함수를 $p_i(x)$라 표시한다. 각 차원상의 데이터는 서로 독립적이라고 가정함에 따라, 이하에서 설명할 알고리즘은 각 차원에 대하여 독립적으로 적용될 수 있다. 또한, 상술한 바와 같이 데이터 분포의 균일성은 실제적으로 데이터의 확률 분포 함수는 불규칙하거나 가우시안 함수등과 같은 잘 정의된 함수에 의하여 모델링되지 않을 수 있다. 본 발명에서는 데이터 분포의 변동을 견디기 위하여 가우시안 혼합 함수를 사용함으로써 일차원 데이터의 확률 분포 함수를 모델링한다.

<31> 먼저, 일차원 신호의 확률 분포 함수 $p(x)$를

<32> 【수학식 1】

$$p(x)=\sum_{j=1}^{M}p(x\mid j)P(j)$$

<33> 과 같이 나타낸다고 가정한다. 여기서, $p(x\mid j)$는,

<34> 【수학식 2】

$$p(x\mid j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\}$$

<35> 과 같이 정의된다. 여기서, 계수 $P(j)$는 혼합 변수라고 칭하며, $0\le P(j)\le 1$의 조건과,

<36> 【수학식 3】

$$\sum_{j=1}^{M}P(j)=1$$

<37> 의 조건을 만족한다. 이로써, 본 발명에 따르면 확률 분포 함수를 가우시안 함수의 가중된 합을 사용하여 정의한다. 이제, 확률 분포 함수를 추정하는 작업은 변수 추정하는 작업으로 귀착된다. 여기서, 추정하여야 하는 변수들은 $j=1,...,M$이고, $l=1,...,N$이며, $v[l]$는 주어진 데이터 세트라 할 때,

<38> 【수학식 4】

$$\Phi(\phi_1,\ldots,\phi_M)=\prod_{l=0}^{N}p(v[l]\mid(\phi_1,\ldots,\phi_M))$$

<39> 를 최대로 하는 Φ 를 구하는 것이다.

<40> 최대화 추정 알고리즘(Expectation-Maximization algorithm: 이하 EM 알고리즘이라 칭함)을 사용하여 변수들을 구한다. 상기 알고리즘에 따르면, 추정을 위한 입력으로써 주어진 N 데이터가 사용되고, 매번 주어진 모든 N 데이터를 사용하여 반복적으로 변수들을 추정한다.

<41> t 를 반복 횟수라 하면, 추정된 변수들은,

<42> 【수학식 5】

$$\mu_j^{t+1}=\frac{\sum_{l=1}^{N}p(j\mid v[l])^t v[l]}{\sum_{l=1}^{N}p(j\mid v[l])^t}$$

<43> 【수학식 6】

$$(\sigma_j^2)^{t+1}=\frac{\sum_{l=1}^{N}p(j\mid v[l])^t(v[l]-\mu_j^t)^2}{\sum_{l=1}^{N}p(j\mid v[l])^t}$$

<44> 【수학식 7】

$$P(j)^{t+1}=\frac{1}{N}\sum_{l=1}^{N}p(j\mid v[l])^t$$

<45> 하지만, 데이터의 분포가 가우시안 함수의 일부로써 그루핑되지 않는 특이값(singular value)를 가지는 경우에는 추정이 오류를 야기시킬 수 있다.

<46> 만일 추정의 오류가 이러한 값을 정확하게 캡쳐하지 못하여 발생하는 경우에는 일정한 μ을 어떤 값으로 근접시켜 대응되는 항 σ^2이 0으로 수렴되도록 하여야 한다. 이러한 특이성 문제를 피하기 위하여, 매우 작은 값을 추정 분산을 위한 하위 경계값(lower bound)로써 설정한다.

<47> 가우시안 혼합함수의 변수들을 추정하기 위하여 EM 알고리즘을 사용하는 것의 효율성을 설명하기 위하여 도 3a에는 특징 벡터 데이터 공간내의 특징 벡터 데이터의 분포를 나타내는 히스토그램을 도시하였으며, 도 3b에는 상기 히스토그램에 대한 추정된 확률 분포 함수의 그래프를 나타내었다. 도 3a와 같이 특징 벡터 데이터의 분포가 불규칙하고, 모델링 함수로써 가우시안 혼합 함수를 사용하고 변수들을 추정하기 위하여 EM 알고리즘을 사용하여 간단한 함수로 모델링될 수 없는 경우임에도 불구하고 도 3b에 나타낸 바와 같이 확률 분포 함수가 잘 모델링될 수 있음을 알 수 있다.

<48> 한편, 상기 수학식 5, 6, 및 7을 사용하여 주어진 N 데이터에 대하여 변수들을 추정할 수 있다. 큰 데이터베이스의 경우, N은 일반적으로 엘리먼트의 총 수에 비하여 적은 부분에 불과하다. 실제적인 데이터베이스 응용에서는 일정 지점에서의 추정을 갱신할 필요가 있다. 예를들어, 보다 양호한 추정을 위하여 보다 큰 데이터 부분을 사용하기를 희망하는 경우가 있을 수 있다. 또는, 데이터베이스가 일정하지 않은 경우에, 데이터의 통계적인 특성이 변화하기 때문에 확률 분포 함수를 다시 추정하여야 한다. 위의 두 경우에서 이전의 추정에 대한 '기록(memory)'을 완전히 지울 필요는 없다. 변수 추정의 관점에서 볼 때, 데이터 집합이 변화할 때는 추정된 확률 분포 함수의 변화를 추적하는 것에 전략이 집중되어야 한다. 이러한 목적을 위하여, 본 발명에서는 추정을 순차적으로 갱신할 수 있는 알고리즘이 제안된다.

<49> N 데이터 $v[I]$을 사용하여 추정된 변수 집합 $\{P(j)^N, \mu_j^N, (\sigma_j^2)^N\}$이 주어질 경우, 새로운 데이터 $v[N+1]$가 들어올 때 갱신된 변수 집합은,

<50> 【수학식 8】

$$\mu_j^{N+1}=\mu_j^N+\theta_j^{N+1}(v[N+1]-\mu_j^N)$$

<51> 과 같이 계산될 수 있다.

<52> 【수학식 9】

$$(\sigma_j^2)^{N+1}=(\sigma_j^2)^N+\theta_j^{N+1}[(v[N+1]-\mu_j^N)^2-(\sigma_j^2)^N]$$

<53> 【수학식 10】

$$P(j)^{N+1}=P(j)^N+\frac{1}{N+1}(p(j \mid v[N+1])-P(j)^N)$$

<54> 상기 수학식 8과 9에서,

<55> 【수학식 11】

$$(\theta_j^{N+1})^{-1}=\frac{p(j \mid v[N])}{p(j \mid v[N+1])}(\theta_j^N)^{-1}+1$$

<56> 의 관계가 성립한다.

<57> 온라인 추정을 사용한 추적 성능을 평가하기 위하여 합성 데이터(synthetic data) 집합에 대한 실험을 수행하였다. 도 4a에는 데이터 집합들의 특징 벡터값들을 나타내었다. 도 4a를 참조하면, 데이터 집합은 5,000개의 엘리먼트들을 포함한다. 도 4b에는 도 4a의 데이터 집합에 대한 히스토그램의 계산 결과를 나타내었다. 각 개별 엘리먼트는 추정을 위하여 순차적으로 더해진다. 이제, 수학식 8, 9, 및 10에 따라 변수들이 계산된다. 다음으로, 일정량의 엘리먼트들이 추정에 사용되었을때 추정된 변수들로부터 확률 분포 함수를 구성한다.

<58> 도 4c, 도 4d, 및 도 4e에는 추정에 사용된 엘리먼트들의 양이 각각 1700, 3400, 및 5000일 때의 추정된 확률 분포 함수를 나타내었다. 도 4c, 도 4d, 및 도 4e를 참조하면, 입력 데이터의 분포가 변화할 때, 온라인 추정은 아주 잘 추적하고 있음을 알 수 있

다. 여기서, 온라인 추정의 효율성은 데이터가 입력으로써 선택되는 방식에 부분적으로 의존한다는 것을 주목할 필요가 있다.

<59> 예를들어, 도 4a에 도시된 데이터의 확률 분포 함수를 추정하고자 할 경우에는 데이터들이 인덱싱된 것과 동일한 순서로 상기 데이터가 선택되었다면 도 4e에 도시한 바와 같은 하나의 추정된 확률 분포 함수를 얻게된다. 즉, 이상적으로는 데이터가 편중되지 않게 선택되어야 함을 알 수 있다.

<60> 이제, 추정된 확률 분포 함수를 $\hat{p}(x)$이라 한다. 비선형 양자화의 목적은 각 그리드에 의하여 커버되는 면적이 동일하도록 확률 분포 함수를 복수 개의 그리드들을 사용하여 분할하는 것이다. 경계점들을 $c[l]$로써 나타낸다고 하면, 상기 경계점들은,

<61> 【수학식 12】

$$\int_{c[l]}^{c[l+1]} \hat{p}(x)dx = \frac{1}{2^b} \int_{c[0]}^{c[2^b]} \hat{p}(x)dx$$

<62> 의 조건(criterion)을 만족하여야 한다. 이러한 조건을 사용하여 추정된 확률 분포 함수의 단일 통과 스켄(one pass scan)으로부터 경계점들을 결정하는 것이 가능하다. 예를들어, 모든 N 점들을 2^b 클러스터들로 응집시킴으로써 수학식 4에서 각 차원의 경계점들을 결정한다. 또한, 상기 수학식 12를 사용하여 계산상 매우 효율적으로 경계점들을 구할 수 있을 뿐만 아니라 차이 측정(distance measure)의 종속성을 회피할 수 있다.

<63> 상기와 같은 방법에 따르면, 확률 분포 함수를 갱신할 수 있는 능력이 있다. 이러한 특징은 균일하지 않은 데이터베이스의 경우에는 만족할만 한 인덱싱을 유지함에 있어서 매우 중요하다. 즉, 이전의 추정이 갱신된 추정과 부합하지 않는 모든 경우에 근사화는 갱신될 필요가 있다. 이러한 이유로, 언제 확률 분포 함수의 추정의 변화를 기초로

근사화를 갱신할 것인지를 결정하기 위한 측정이 요구된다. 또한, 근사화를 구성하는 확률 분포 함수를 사용하는 병렬적 스킴으로써 근사화를 갱신하기 위한 측정은 각 차원에 대하여 정의될 수 있다. 이전 확률 분포 함수와 갱신된 확률 분포 함수를 각각 $\hat{p}_{old}(x)$ 과 $\hat{p}_{new}(x)$라 할 때, 확률 분포 함수의 변화의 측정은,

<64> 【수학식 13】

$$\rho = \frac{\int (\hat{p}_{old}(x) - \hat{p}_{new}(x))^2 dx}{\int \hat{p}_{old}(x)^2 dx}$$

<65> 과 같이 정의될 수 있다. 여기서, ρ가 소정의 임계값보다 클 때 그 차원에 대한 근사화가 갱신된다.

<66> 34,698개의 항공 사진 영상들을 포함하는 영상 데이터베이스에 대하여 평가를 위한 모의 실험을 실행하였다. 먼저, 소정의 텍스쳐 추출 방법을 사용하여 각 영상들에 대하여 영상의 텍스쳐 특징을 기술하는 48 차원의 특징 벡터를 추출한다. 추출된 특징 벡터를 기초로 하는 전체 데이터 집합으로부터 확률 분포 함수가 추정된다. 도 5a 및 도 5b에는 종래의 인덱싱 방법과 본 발명의 인덱싱 방법을 사용하여 제1 단계 필터링 및 제2 단계 필터링에서 방문한 특징 벡터의 수를 비교 도시한 그래프를 나타내었다. 도 5a에서, 그래프(502)는 적응적으로 VA 파일을 구성하는 본원 발명의 인덱싱 방법을 사용할 때의 제1 단계 필터링에서 방문한 특징 벡터의 수를 나타내며, 그래프(504)는 고정된 VA 파일을 사용하는 종래의 인덱싱 방법을 사용할 때의 제1 단계 필터링에서 방문한 특징 벡터의 수를 나타낸다. 또한, 제1 단계 필터링에서 방문한 특징 벡터의 수를 수직축으로써 N1으로 표시하였다.

<67> 또한, 도 5b에서 그래프(512)는 적응적으로 VA 파일을 구성하는 본원 발명의 인덱

싱 방법을 사용할 때의 제2 단계 필터링에서 방문한 특징 벡터의 수를 나타내며, 그래프(514)는 고정된 VA 파일을 사용하는 종래의 인덱싱 방법을 사용할 때의 제2 단계 필터링에서 방문한 특징 벡터의 수를 나타낸다. 또한, 제2 단계 필터링에서 방문한 특징 벡터의 수를 수직축으로써 N2으로 표시하였다. 그래프(502, 504)와 그래프(512, 514)를 비교하면, 고정된 VA 파일을 사용하는 종래의 인덱싱 방법을 사용할 때의 제1 단계 필터링 및 제2 단계 필터링에서 방문한 특징 벡터의 수 보다 적응적으로 VA 파일을 구성하는 본 원 발명의 인덱싱 방법을 사용할 때의 제1 단계 필터링 및 제2 단계 필터링에서 방문한 특징 벡터의 수가 상당히 크다는 것을 알 수 있다.

<68> 상기와 같은 본 발명에 따른 인덱싱 방법은 개인용 또는 서버급의 컴퓨터내에서 실행되는 프로그램으로 작성 가능하다. 상기 프로그램을 구성하는 프로그램 코드들 및 코드 세그멘트들은 당해 분야의 컴퓨터 프로그래머들에 의하여 용이하게 추론될 수 있다. 또한, 상기 프로그램은 컴퓨터 독취 가능 기록 매체에 저장될 수 있다. 상기 기록 매체는 자기기록매체, 광기록 매체, 및 전파 매체를 포함한다.

【발명의 효과】

<69> 상술한 바와 같이 본 발명에 의한 특징 벡터 데이터 공간의 인덱싱 방법은 특징 벡터들이 균일하게 분포하지 않는 차수가 높은 벡터 공간내에서 특징 벡터 데이터 공간을 효율적으로 인덱싱한다. 또한, 상기 특징 벡터 데이터 공간의 인덱싱 방법은 새로운 특징 벡터 데이터가 추가되었을때 인덱싱의 업그레이드가 용이하다.

【특허청구범위】

【청구항 1】

특징 벡터 데이터 공간내에서 특징 벡터를 인덱싱하는 방법에 있어서,

(a) 특징 벡터 데이터 공간내의 특징 벡터 데이터의 통계적 분포를 기초로 특징 벡터들을 적응적으로 근사화함으로써 특징 벡터 데이터 공간을 인덱싱하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 2】

제1항에 있어서, 상기 (a) 단계는,

(a-1) 특징 벡터 데이터 공간내에서 특징 벡터 데이터의 통계적인 분포를 측정하는 단계;

(a-2) 상기 통계적인 분포를 사용하여 데이터의 주변 분포를 추정하는 단계;

(a-3) 상기 추정된 분포를 그 데이터가 어떠한 각 그리드에 놓이는 확률이 균등하게 되는 복수 개의 그리드들로 분할하는 단계; 및

(a-4) 분할된 그리드들을 사용하여 특징 벡터 데이터 공간을 인덱싱하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 3】

제2항에 있어서, 상기 (a-4) 단계 이전에,

새로운 데이터가 들어오면 이전 확률 분포 함수와 갱신된 확률 분포 함수를 기초로 그리드를 갱신하는 단계;를 더 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 4】

제2항에 있어서, 상기 (a-4) 단계는,

VA(vector approximation: 벡터 근사화) 파일을 사용하여 인덱싱하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 5】

제2항에 있어서, 상기 복수 개의 그리드의 수는 그 차원에 할당된 주어진 비트수에 의하여 결정된 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 6】

제2항에 있어서, 상기 (a-2) 단계는,

(a-2-1) 확률 분포 함수를 소정의 분포 함수의 가중된 합을 사용하여 정의하는 단계; 및

(a-2-2) 상기 (a-2-1) 단계에서 정의된 확률 분포 함수를 사용하여 소정의 변수들을 추정함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 7】

제6항에 있어서, 상기 (a-2-2) 단계는,

상기 (a-2-1) 단계에서 정의된 확률 분포 함수를 사용하여 최대화 추정 알고리즘을 기초로, N은 소정의 양의 수, 매번 주어진 모든 N 데이터를 사용하여 반복적으로 소정의 변수들을 추정함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 8】

제6항에 있어서, 상기 소정의 분포 함수는,

가우시안 함수인 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 9】

제6항에 있어서, 상기 (a-2-1) 단계는,

$p(x|j)$는 $p(x|j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\}$로서 정의되고, 혼합 변수라고 칭해지는 계수 $P(j)$는 $0\le P(j)\le 1$의 조건과, $\sum_{j=1}^{M}P(j)=1$의 조건을 만족한다고 할 때, 일차원 신호의 확률 분포 함수 $p(x)$를 $p(x)=\sum_{j=1}^{N}p(x|j)P(j)$라고 가정하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 10】

제6항에 있어서, 상기 (a-2-2) 단계는,

추정하여야 하는 변수들은 $j=1,\ldots,M$이고, $l=1,\ldots,N$이며, $v[l]$는 주어진 데이터 세트라 할 때, $\Phi(\phi_1,\ldots,\phi_M)=\prod_{l=0}^{N}p(v[l]\mid(\phi_1,\ldots,\phi_M))$를 최대로 하는 Φ를 구함으로써 추정된 확률 분포 함수를 구하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 11】

제10항에 있어서, 상기 (a-2-2) 단계는,

t는 반복 횟수를 나타내는 양의 수라 할 때, $\mu_j^{t+1}=\frac{\sum_{l=1}^{N}p(j\mid v[l])^t v[l]}{\sum_{l=1}^{N}p(j\mid v[l])^t}$,

$(\sigma_j^2)^{t+1}=\frac{\sum_{l=1}^{N}p(J\mid v[l])^t(v[l]-\mu_j^t)^2}{\sum_{l=1}^{N}p(J\mid v[l])^t}$, $P(J)^{t+1}=\frac{1}{N}\sum_{l=1}^{N}p(J\mid v[l])^t$에 따라 추정된 변수들을 구하는 단계;를 더 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 12】

제5항에 있어서, 상기 (a-2-2) 단계는,

N 데이터 $v[l]$을 사용하여 추정된 변수 집합 $\{P(J)^N,\mu_j^N,(\sigma_j^2)^N\}$이 주어질 때, 새로운 데이터 $v[N+1]$가 들어오면 $\mu_j^{N+1}=\mu_j^N+\theta_j^{N+1}(v[N+1]-\mu_j^N)$, $(\sigma_j^2)^{N+1}=(\sigma_j^2)^N+\theta_j^{N+1}[(v[N+1]-\mu_j^N)^2-(\sigma_j^2)^N]$, $P(J)^{N+1}=P(J)^N+\frac{1}{N+1}(p(J\mid v[N+1])-P(J)^N)$, 및 $(\theta_j^{N+1})^{-1}=\frac{p(J\mid v[N])}{p(J\mid v[N+1])}(\theta_j^N)^{-1}+1$을 계산함으로써 갱신된 변수 집합을 구하는 단계;를 더 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【청구항 13】

제11항에 있어서, 상기 (a-2-2) 단계는,

이전 확률 분포 함수와 갱신된 확률 분포 함수를 각각 $\hat{p}_{old}(x)$과 $\hat{p}_{new}(x)$라 할 때, 각 차원에 대하여 $\rho=\frac{\int(\hat{p}_{old}(x)-\hat{p}_{new}(x))^2dx}{\int\hat{p}_{old}(x)^2dx}$으로 정의된 확률 분포 함수의 변화를 측정하는 단계; 및

ρ가 소정의 임계값보다 크면 그 차원에 대하여 근사화를 갱신하는 단계;를 더 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

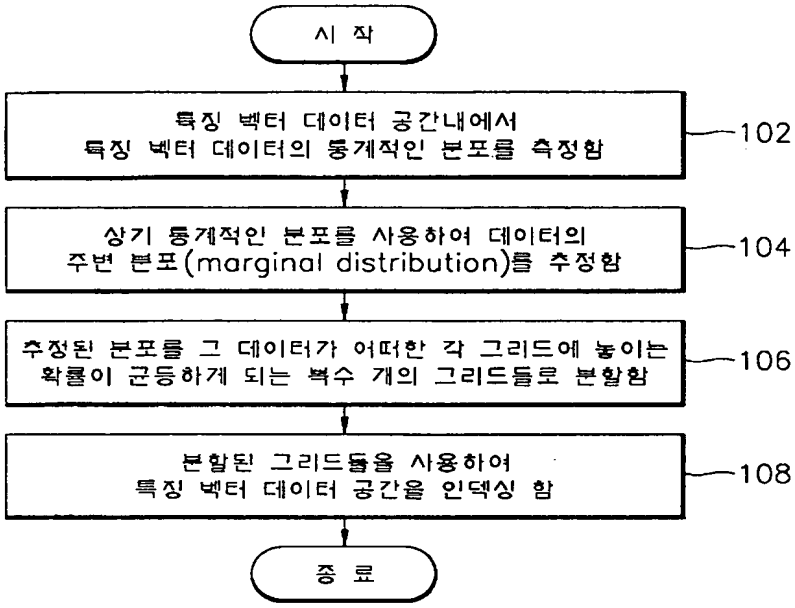
【청구항 14】

제12항에 있어서, 상기 (a-3) 단계는,

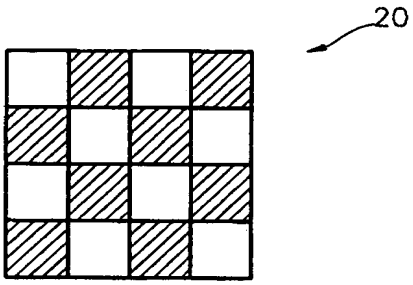
추정된 확률 분포 함수를 $\hat{p}(x)$라 할 때, $\int_{c[l]}^{c[l+1]} \hat{p}(x)dx = \frac{1}{2^b} \int_{c[0]}^{c[2^b]} \hat{p}(x)dx$의 조건 (criterion)을 만족하는 경계점들 $c[l]$에 의하여 결정되는 복수 개의 그리드들을 사용하여 각 그리드에 의하여 커버되는 면적이 동일하도록 확률 분포 함수를 분할하는 단계;를 포함하는 것을 특징으로 하는 특징 벡터 데이터 공간의 인덱싱 방법.

【도면】

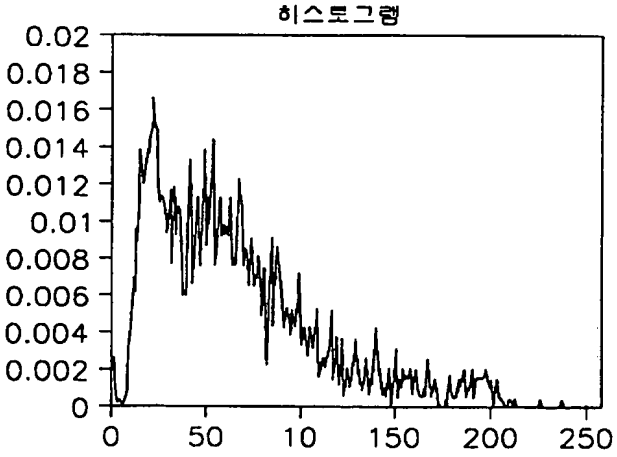
【도 1】


시 작
특징 벡터 데이터 공간내에서
특징 벡터 데이터의 통계적인 분포를 측정함
102
상기 통계적인 분포를 사용하여 데이터의
주변 분포(marginal distribution)를 추정함
104
추정된 분포를 그 데이터가 어떠한 각 그리드에 놓이는
확률이 균등하게 되는 복수 개의 그리드들로 분할함
106
분할된 그리드들을 사용하여
특징 벡터 데이터 공간을 인덱싱 함
108
종 료

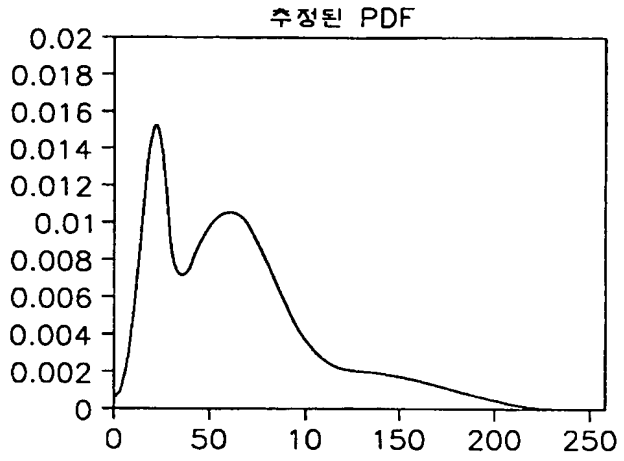
【도 2】


20

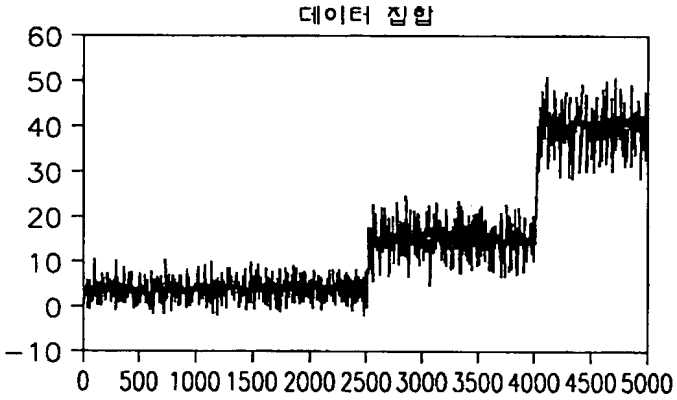
【도 3a】


히스토그램
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
10
150
200
250

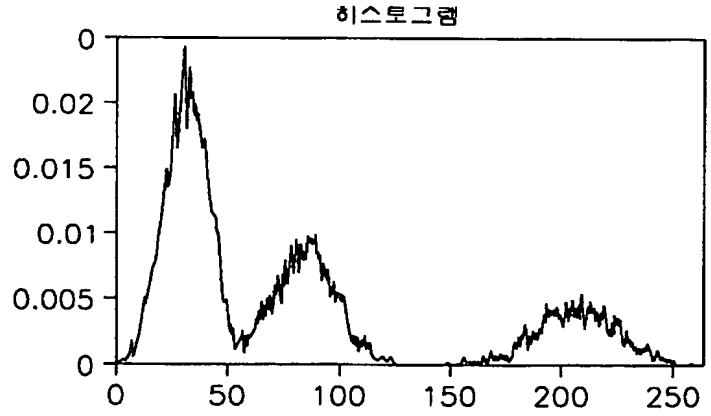
【도 3b】


추정된 PDF
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0 50 10 150 200 250

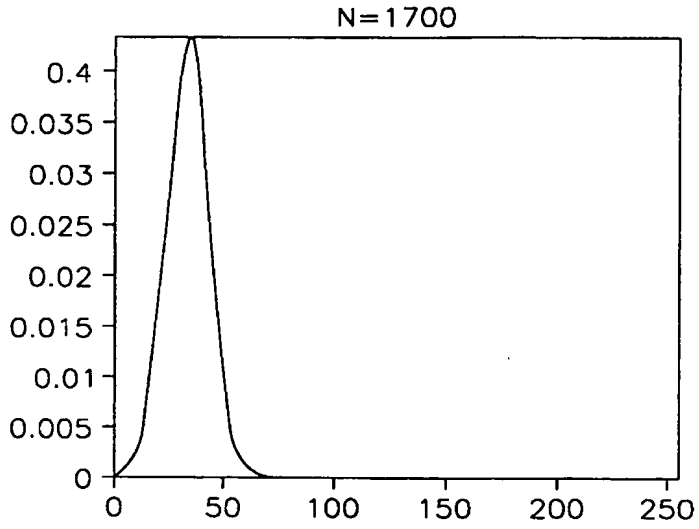
【도 4a】


데이터 집합
60
50
40
30
20
10
0
−10
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000

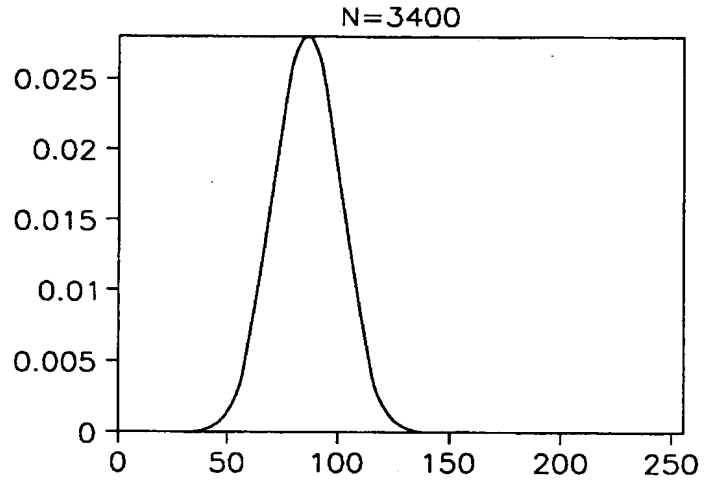
【도 4b】


히스토그램
0
0.02
0.015
0.01
0.005
0
0 50 100 150 200 250

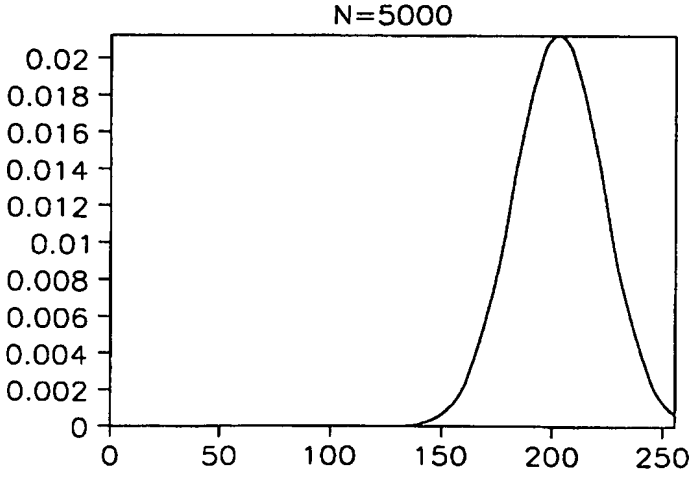
【도 4c】


N=1700
0.4
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
50
100
150
200
250

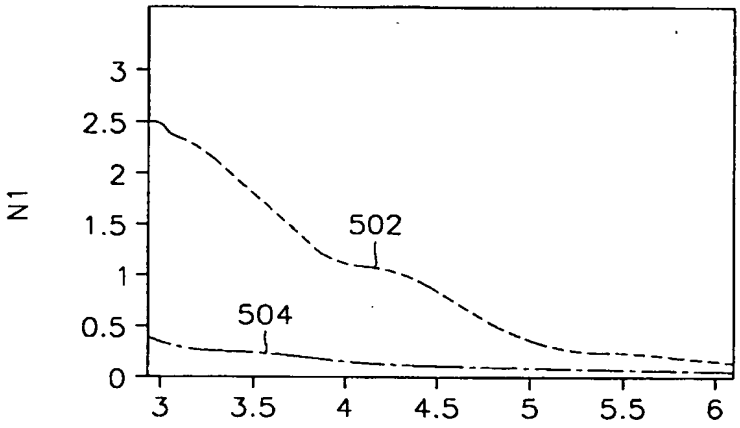
【도 4d】


N=3400
0.025
0.02
0.015
0.01
0.005
0
0
50
100
150
200
250

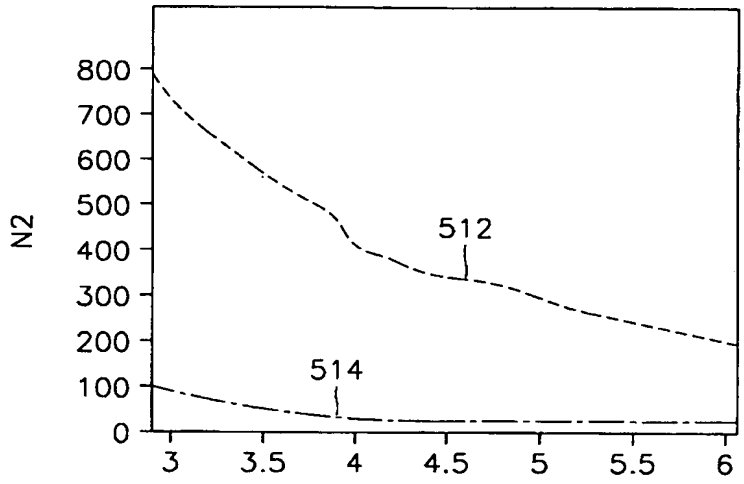
【도 4e】


N=5000
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250

【도 5a】


N1
3
2.5
2
1.5
1
0.5
0
502
504
3
3.5
4
4.5
5
5.5
6

【도 5b】


N2
800
700
600
500
400
300
200
100
0
512
514
3
3.5
4
4.5
5
5.5
6